Journal of Advances in Mathematical & Computational Sciences
An International Pan-African Multidisciplinary Journal of the SMART Research Group
International Centre for IT & Development (ICITD) USA

Available online at https://www.isteams.net/ mathematics-computationaljournal.info
CrossREF Member Listing - https://www.crossref.org/06members/50go-live.html

Exploring the Efficacy of Machine Learning Techniques for Predicting Students' Academic Performances

[1]Agwi, C. U. & [2]Akpojaro, J.
Department of Mathematics and Computer Science
Faculty of Basic and Applied Sciences
University of Africa
Toru-Orua, Bayelsa State
Ucheworld2015@gmail.com, jackson.akpojaro@uat.edu.ng

ABSTRACT

Predicting students' academic success in tertiary institutions in terms of determining accurately whether a student will be a top 10 in his class, an average student or a dropout is a significant issue in higher education. Therefore, to predict the behavior of a learner, many data mining techniques are used, such as clustering, classification, and regression. In this paper, students' academic performance prediction model and new features are introduced that have a great influence on student's overall academic achievement. Twenty features were considered from the dataset obtained from about 200 students using questionnaire and information from the institutions database. The Naive Bayes classifier was used and the experiment was carried out in a WEKA implementation workbench. The attributes were first analyzed for relevance, the analysis showed that a student's first semester grade point average (GPA) has the highest relevance followed by friends' study affinity, learning facilities, private home lesson, gender and available transport facilities. One hundred and sixty-nine (169) students' records were used for the analysis, the model recorded 162 instances that were correctly classified which translates to 95.86% accuracy and incorrectly predicted 4.142%. The experiment showed an excellent performance from the Naive Bayes classifier in predicting the dataset correctly

Keywords: Educational Data Mining, Students' Performance Management, Tertiary Institutions, Classification, Naive Bayes Classifier.

Agwi, C.U. & Akpojaro, J. (2024): Exploring the Efficacy of Machine Learning Techniques for Predicting Students' Academic Performances. Journal of Advances in Mathematical & Computational Science. Vol. 12, No. 1. Pp 11-24. Available online at www.isteams.net/mathematics-computationaljournal. dx.doi.org/10.22624/AIMS/MATHS/V12N2P2

1. INTRODUCTION

Predicting students' academic success in tertiary institutions in terms of determining accurately whether a student will be an academic success, a dropout, or an average performer is a significant issue in higher education. The ability to predict students' performance in an educational environment is very important in students' performance management. With the development of portals, numerous institutions now maintain their educational data in electronic format and this has increased drastically over the years (Danial *et al*. 2019; Aysha *et al.*, 2018 & Nawal *et al.,* 2017). The data in the databases of various educational institutions consists of both historical and operational data, and organizing this huge amount of data to determine the relationships among its variables is a very complex task (Rahila & Suriadi, 2017). Leveraging on data mining techniques with machine learning algorithm, hidden patterns and relationships that are helpful in decision making can be uncovered from such a vast amount of data repositories in today's educational environments. Educational data mining and its application has increasingly become a promising area that offers tools and techniques that can help in attaining this objective.

Students' academic performance depends on several factors that may include personal, social, psychological, and other environmental variables (Sirwan & Ardalan, 2019). The requirements possessed and presented by students can only determine the suitability (or otherwise) of students for admission at the point of entry. However, the student's efficacy can only be established or dispelled by the subsequent performance of the student during the study. The existing educational assessment system (semester or sessional coursework and examination-based scoring system) does not have predictive mechanisms to predict the weak student so that the teachers can be informed early enough for closer attention. Several research studies carried out by scholars have shown that it is inadequate in predicting which student is likely to pass or fail, dropout, or advance (Sirwan & Ardalan, 2019; Kalpesh *et al*., 2017 & Amjad, 2016).

The prediction of students' performance with high accuracy using well thought out parameters and variables that affect student academic performance from the Nigerian educational context will be very beneficial to institutional management. This can help institutions identify the students with low academic impetus early enough so that the identified students can be guided closely by the teacher to improve their performance. It is in connection with this that this study is being carried out. The paper explores the efficacies of predicting students' academic performance using a machine learning model with several variables capable of influencing the academic performance of students.

2. LITERATURE REVIEW

Student academic performance refers to a measure of students' academic achievements in a given period of time, considering different factors that may affect this measure (Romero & Ventura, 2010; Quadri & Kalyankar, 2010). Student academic performance prediction involves investigating and analyzing the past and present values of variables that describe a student, in order to provide an estimate of a certain future output for that same student, such as the student's marks and student's attendance. The prediction process involves estimating interesting (and sometimes unexpected) connections between the output (performance in this case) and various inputs using machine learning algorithms. Many scholarly works have been carried out by some scholars in this research area. Ganorkar *et al*. (2021) reviewed the work done by various researchers on analysis and prediction of students' data for identifying important attributes.

The result showed that the development of algorithm and inferring the data is now an integral part of the data science, used for analyzing and predicting the result of the large form of database using data mining techniques. Usman *et al*. (2020) proposed a classification technique for analyzing and evaluating student academic performance using a decision tree model and in a Rapid Miner tool with data collected over a period of two (2) years. Hasan *et al.* (2020) proposed a system to predict student's overall performance at the end of the semester using video learning analytics and data mining techniques. The results showed that Random Forest accurately predicted successful students at the end of the class with an accuracy of 88.3% with an equal width and information gain ratio.

Mohammed & Nawzat (2019) carried out a research in which they classified and predicted students' performance using classification algorithms in the final exam. Four classification algorithms, which are Decision Tree C4.5, Random Forest, Support Vector Machine (SVM) and Naive Bayes, were used in this research in order to classify and predict the students' performance. Furthermore, their study was aimed at improving the Decision Tree C4.5 algorithm by adding a grid search function in order to improve prediction accuracy in classifying and predicting the students' performance.

Annisa & Harwati (2019) undertook a study in which they tried to implement two popular data mining – clustering and classification analysis to predict student's performance. Saheed *et al.* (2018) performed an experiment in order to presents a method to predict student performance using Iterative dichotomized 3 (ID3), C4.5 and Classification and Regression tree (CART). The experiment was performed on Waikato Environment for Knowledge Analysis (Weka). Experiment results revealed that C4.5 outperforms other classifiers and requires reasonable amount of time to build the model. Nawal *et al.* (2017) undertook a study on predicting student academic performance in KSA using data mining techniques with the specific objective of finding out if there are any patterns in the available data (student and courses records) that could be useful for predicting students' performance.

The study involved a sample of 150 students collected from Najran University students in Saudi Arabia. It is evident that studies have been undertaken in the field of students' academic performance prediction using various techniques. However, there exists a very significant gap as could be seen from the reviewed literature which indicates that most of these studies were plagued with insufficient parameters, variables or attributes to provide adequate and efficient prediction in the education sector.

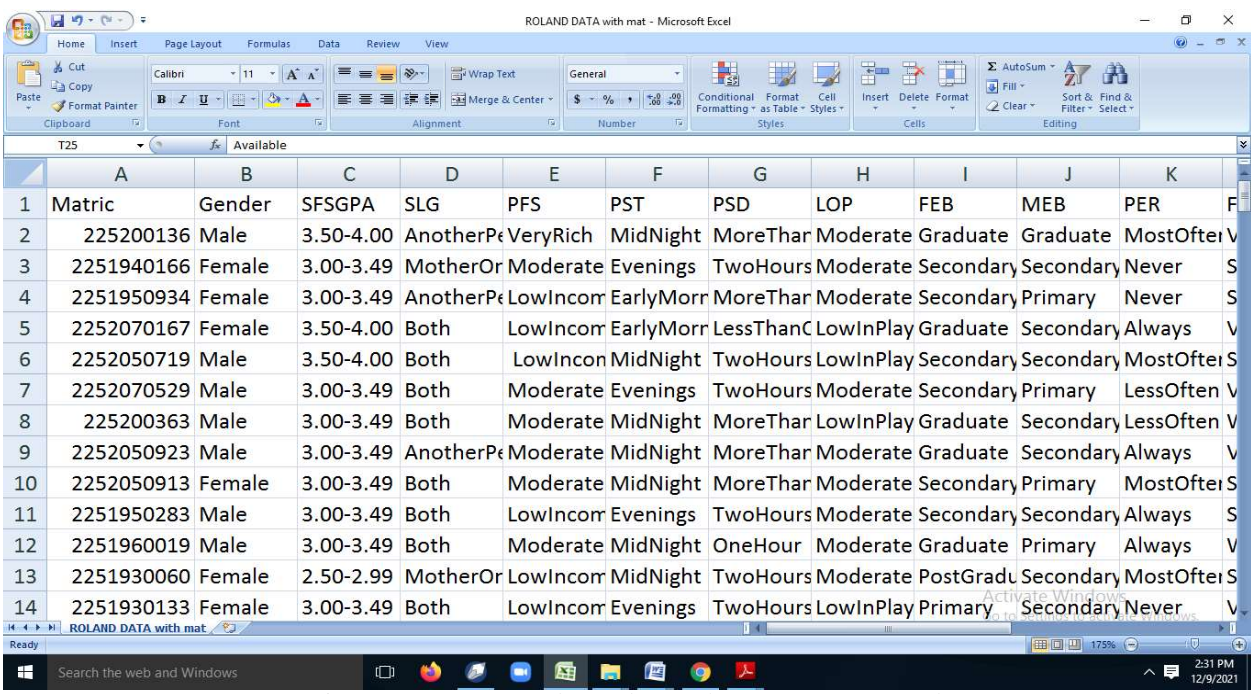
3. METHODOLOGY

3.1 Dataset Preparation

Primary data were collected from polytechnic students using questionnaire as the data gathering tool. The questionnaire consisted of questions related to several personal, socio-economic, psychological and school related variables that were considered capable of affecting students' academic performance. The questionnaire consisted 20 questions administered to 200 students. The secondary data such as mark details were collected from the institutions. All the predictor and response variables which were derived from the questionnaire were properly tabulated. The domain values for some of the identified variables were: gender (student sex), school life guardian (SLG).

This determines the persons who took care of the student during his school years, parents financial status (PFS), personal study time (PST), personal study duration (PSD), level of play (LOP), father's educational background (FEB), mother's educational background (MEB), parents enforce reading (PER), friends study affinity (FSA), secondary school type (SST), private home lesson (PHL), Food availability (FA), health status (HS), available transportation facility (ATF), classroom condition (CC), senior secondary school performance (SSP), learning facilities (LF), and student first semester GPA (SFSGPA).

The data were carefully entered into the Excel spreadsheet first, subsequently converted to Comma Separated Version (CSV) format and to Attribute-Relation File Format (ARFF) recognizable by WEKA application. Figure 1 shows the snap shot of the dataset in Excel format.



	A	B	C	D	E	F	G	H	I	J	K	
1	Matric	Gender	SFSGPA	SLG	PFS	PST	PSD	LOP	FEB	MEB	PER	F
2	225200136	Male	3.50-4.00	AnotherPe	VeryRich	MidNight	MoreThar	Moderate	Graduate	Graduate	MostOftei	V
3	2251940166	Female	3.00-3.49	MotherOr	Moderate	Evenings	TwoHours	Moderate	Secondary	Secondary	Never	S
4	2251950934	Female	3.00-3.49	AnotherPe	LowIncom	EarlyMorr	MoreThar	Moderate	Secondary	Primary	Never	S
5	2252070167	Female	3.50-4.00	Both	LowIncom	EarlyMorr	LessThanC	LowInPlay	Graduate	Secondary	Always	V
6	2252050719	Male	3.50-4.00	Both	LowIncon	MidNight	TwoHours	LowInPlay	Secondary	Secondary	MostOftei	S
7	2252070529	Male	3.00-3.49	Both	Moderate	Evenings	TwoHours	Moderate	Secondary	Primary	LessOften	V
8	225200363	Male	3.00-3.49	Both	Moderate	MidNight	MoreThar	LowInPlay	Graduate	Secondary	LessOften	V
9	2252050923	Male	3.00-3.49	AnotherPe	Moderate	MidNight	MoreThar	Moderate	Graduate	Secondary	Always	V
10	2252050913	Female	3.00-3.49	Both	Moderate	MidNight	MoreThar	Moderate	Secondary	Primary	MostOftei	S
11	2251950283	Male	3.00-3.49	Both	LowIncom	Evenings	TwoHours	Moderate	Secondary	Secondary	Always	S
12	2251960019	Male	3.00-3.49	Both	Moderate	MidNight	OneHour	Moderate	Graduate	Primary	Always	V
13	2251930060	Female	2.50-2.99	MotherOr	LowIncom	MidNight	TwoHours	Moderate	PostGradu	Secondary	MostOftei	S
14	2251930133	Female	3.00-3.49	Both	LowIncom	Evenings	TwoHours	LowInPlay	Primary	Secondary	Never	V

Figure 1: Implementation Snap Shot of the data in Excel format.

3.2 System Architecture for the Study

The architecture of the system (Figure 2) shows the processes involved in performing the prediction using Naïve Bayes Classifiers (NBC) as the machine learning algorithm. The techniques were immediately applied to the dataset to create the model using the WEKA tool as the implementation environment. The model performance was evaluated using the testing data.

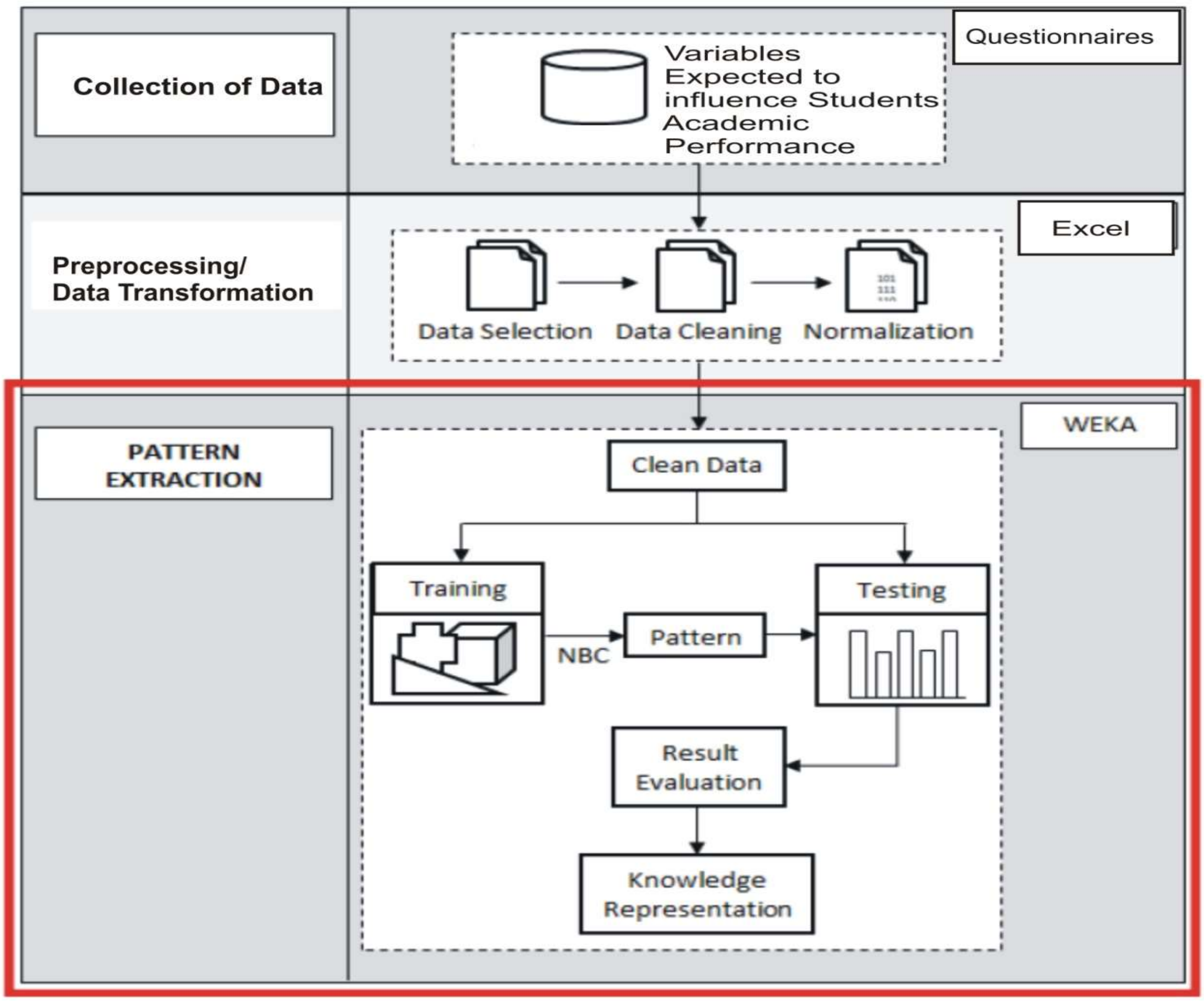


Figure 2. System architecture for student academic performance prediction

3.3 Naïve Bayes Classifier (NBC) Algorithm

Naive Bayes Classifiers are a collection of classification algorithms based on Bayes' Theorem.

It is based on the following assumptions, which include;

i. the features of the data are conditionally independent of each other, given the class label (features independence).
ii. if a feature is continuous, then it is assumed to be normally distributed within each class.
iii. if a feature is discrete, then it is assumed to have a multinomial distribution within each class.
iv. all features are assumed to contribute equally to the prediction of the class label.
v. the data should not contain any missing values.

The formula for Naïve Bayes is given as;

$$p(A\backslash B) = \frac{p(B\backslash A)p(A)}{\sum_{allA} p(B\backslash A)p(A)} \quad \text{---} \quad (1)$$

Where:
$p(A\backslash B)$ is called posterior. This quantity is defined as the probability of the hypothesis given the data.
$p(B\backslash A)$ is called likelihood. It's is the probability of the data given the hypothesis.
$p(A)$ is called prior. It represents our belief about the distribution.
$p(B)$ is defined as the normalizing constant. The $p(B)$ is calculated as;

$$p(B) = \sum_{allA} p(B\backslash A)p(A) \quad \text{---} \quad (2)$$

$p(B)$ ensures that the sum of the posterior over A is equal to 1.

Naive Bayes Classifier offers the following advantages which include; easy to implement and computationally efficient, effective in cases with a large number of features, performs well even with limited training data, performs well in the presence of categorical features and for numerical features data is assumed to come from normal distributions. Naive Bayes Classifier works as follows;

Step 1: the dataset is converted into a frequency table
Step 2: Likelihood table is created by finding the probabilities
Step 3: the posterior probability for each class is computed using equation (1)
Step 4: the class with the highest posterior probability is extracted as the outcome of prediction.

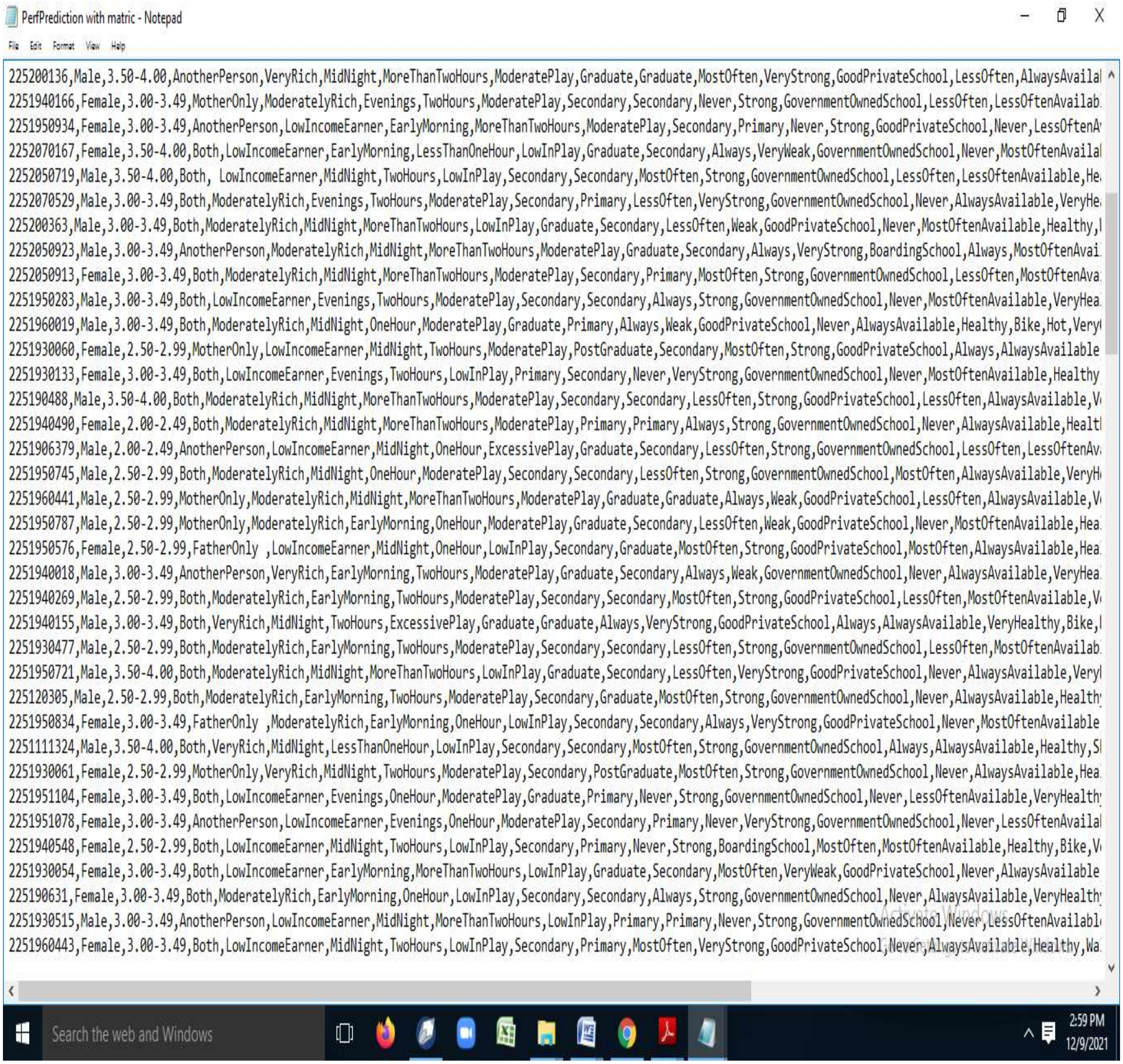
4. IMPLEMENTATION AND ANALYSIS OF RESULTS

The implementation stages include; data transformation, preprocessing, model training and testing.

4.1 Data Transformation

The Data transformation stage consists of three phases which are data selection, data cleaning, and data normalization. This stage is implemented to improve the data quality for mining to obtain more accurate and precise results of the analysis. Twenty parameters were selected for the mining process. In the data cleaning phase, the missing and incomplete value/data were removed.

The last phase was data normalization, where the numerical value such as GPA were transformed into nominal value. For example, the GPA range within 0.00-1.99 is categorized as "Poor", the GPA range within 2.00-2.49 as "Pass", the GPA range within 2.50-2.99 as "Lower Credit" the GPA range within 3-00-3.49 as "Upper Credit" and lastly the GPA range within 3.50-4.00 as "Distinction". In line with this, the data in the Excel format was saved to a comma separated values format and then opened the new file in a note pad text editor as shown in Figure 3.

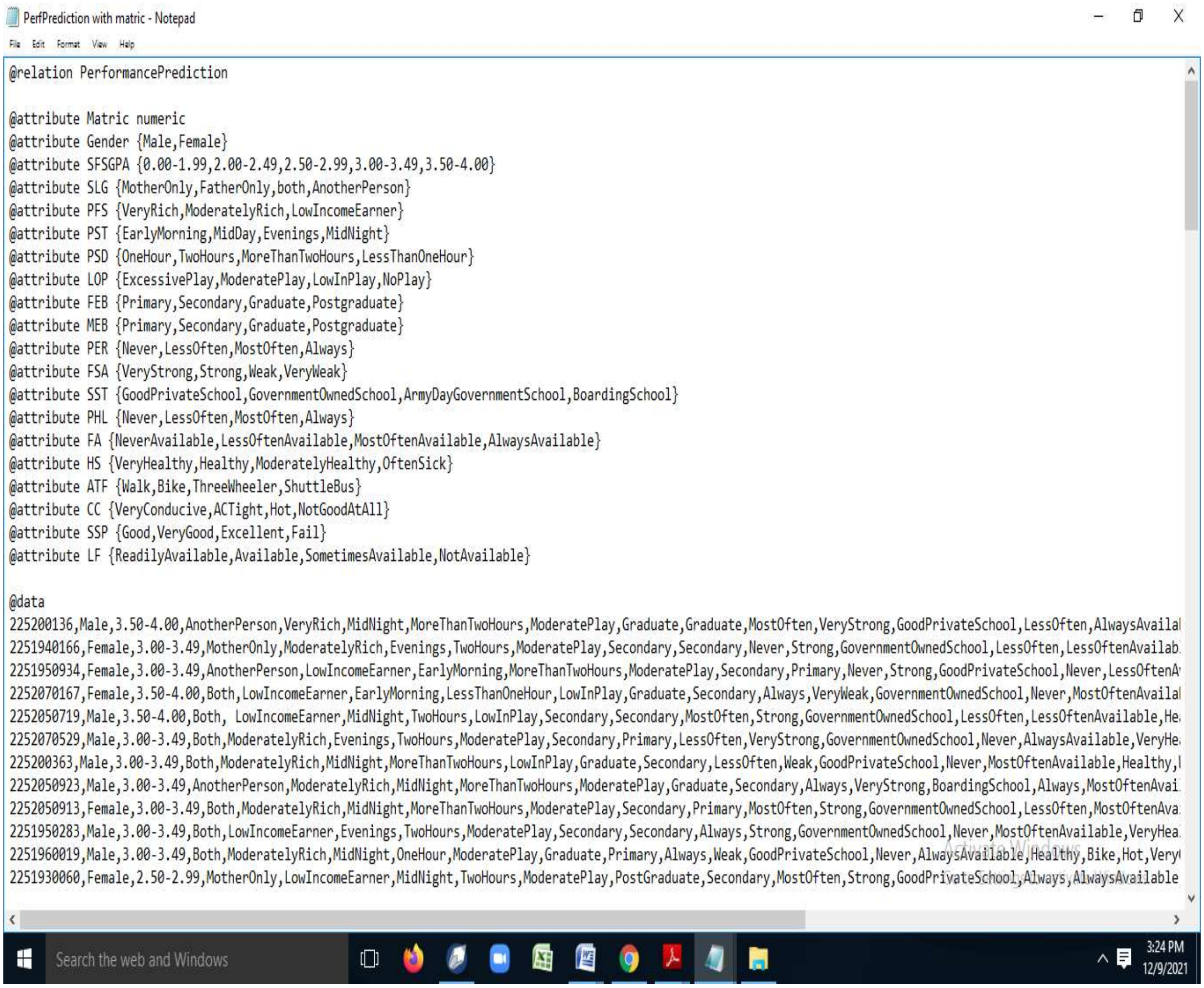

PerfPrediction with matric - Notepad
File Edit Format View Help

```
225200136,Male,3.50-4.00,AnotherPerson,VeryRich,MidNight,MoreThanTwoHours,ModeratePlay,Graduate,Graduate,MostOften,VeryStrong,GoodPrivateSchool,LessOften,AlwaysAvailal
2251940166,Female,3.00-3.49,MotherOnly,ModeratelyRich,Evenings,TwoHours,ModeratePlay,Secondary,Secondary,Never,Strong,GovernmentOwnedSchool,LessOften,LessOftenAvailab
2251950934,Female,3.00-3.49,AnotherPerson,LowIncomeEarner,EarlyMorning,MoreThanTwoHours,ModeratePlay,Secondary,Primary,Never,Strong,GoodPrivateSchool,Never,LessOftenA
2252070167,Female,3.50-4.00,Both,LowIncomeEarner,EarlyMorning,LessThanOneHour,LowInPlay,Graduate,Secondary,Always,VeryWeak,GovernmentOwnedSchool,Never,MostOftenAvailal
2252050719,Male,3.50-4.00,Both, LowIncomeEarner,MidNight,TwoHours,LowInPlay,Secondary,Secondary,MostOften,Strong,GovernmentOwnedSchool,LessOften,LessOftenAvailable,He
2252070529,Male,3.00-3.49,Both,ModeratelyRich,Evenings,TwoHours,ModeratePlay,Secondary,Primary,LessOften,VeryStrong,GovernmentOwnedSchool,Never,AlwaysAvailable,VeryHe
225200363,Male,3.00-3.49,Both,ModeratelyRich,MidNight,MoreThanTwoHours,LowInPlay,Graduate,Secondary,LessOften,Weak,GoodPrivateSchool,Never,MostOftenAvailable,Healthy,
2252050923,Male,3.00-3.49,AnotherPerson,ModeratelyRich,MidNight,MoreThanTwoHours,ModeratePlay,Graduate,Secondary,Always,VeryStrong,BoardingSchool,Always,MostOftenAvai
2252050913,Female,3.00-3.49,Both,ModeratelyRich,MidNight,MoreThanTwoHours,ModeratePlay,Secondary,Primary,MostOften,Strong,GovernmentOwnedSchool,LessOften,MostOftenAva
2251950283,Male,3.00-3.49,Both,LowIncomeEarner,Evenings,TwoHours,ModeratePlay,Secondary,Secondary,Always,Strong,GovernmentOwnedSchool,Never,MostOftenAvailable,VeryHea
2251960019,Male,3.00-3.49,Both,ModeratelyRich,MidNight,OneHour,ModeratePlay,Graduate,Primary,Always,Weak,GoodPrivateSchool,Never,AlwaysAvailable,Healthy,Bike,Hot,Very
2251930060,Female,2.50-2.99,MotherOnly,LowIncomeEarner,MidNight,TwoHours,ModeratePlay,PostGraduate,Secondary,MostOften,Strong,GoodPrivateSchool,Always,AlwaysAvailable
2251930133,Female,3.00-3.49,Both,LowIncomeEarner,Evenings,TwoHours,LowInPlay,Primary,Secondary,Never,VeryStrong,GovernmentOwnedSchool,Never,MostOftenAvailable,Healthy
225190488,Male,3.50-4.00,Both,ModeratelyRich,MidNight,MoreThanTwoHours,ModeratePlay,Secondary,Secondary,LessOften,Strong,GoodPrivateSchool,LessOften,AlwaysAvailable,V
2251940490,Female,2.00-2.49,Both,ModeratelyRich,MidNight,MoreThanTwoHours,ModeratePlay,Primary,Primary,Always,Strong,GovernmentOwnedSchool,Never,AlwaysAvailable,Healt
2251906379,Male,2.00-2.49,AnotherPerson,LowIncomeEarner,MidNight,OneHour,ExcessivePlay,Graduate,Secondary,LessOften,Strong,GovernmentOwnedSchool,LessOften,LessOftenAv
2251950745,Male,2.50-2.99,Both,ModeratelyRich,MidNight,OneHour,ModeratePlay,Secondary,Secondary,LessOften,Strong,GovernmentOwnedSchool,MostOften,AlwaysAvailable,VeryH
2251960441,Male,2.50-2.99,MotherOnly,ModeratelyRich,MidNight,MoreThanTwoHours,ModeratePlay,Graduate,Graduate,Always,Weak,GoodPrivateSchool,LessOften,AlwaysAvailable,V
2251950787,Male,2.50-2.99,MotherOnly,ModeratelyRich,EarlyMorning,OneHour,ModeratePlay,Graduate,Secondary,LessOften,Weak,GoodPrivateSchool,Never,MostOftenAvailable,Hea
2251950576,Female,2.50-2.99,FatherOnly ,LowIncomeEarner,MidNight,OneHour,LowInPlay,Secondary,Graduate,MostOften,Strong,GoodPrivateSchool,MostOften,AlwaysAvailable,Hea
2251940018,Male,3.00-3.49,AnotherPerson,VeryRich,EarlyMorning,TwoHours,ModeratePlay,Graduate,Secondary,Always,Weak,GovernmentOwnedSchool,Never,AlwaysAvailable,VeryHea
2251940269,Male,2.50-2.99,Both,ModeratelyRich,EarlyMorning,TwoHours,ModeratePlay,Secondary,Secondary,MostOften,Strong,GoodPrivateSchool,LessOften,MostOftenAvailable,V
2251940155,Male,3.00-3.49,Both,VeryRich,MidNight,TwoHours,ExcessivePlay,Graduate,Graduate,Always,VeryStrong,GoodPrivateSchool,Always,AlwaysAvailable,VeryHealthy,Bike,
2251930477,Male,2.50-2.99,Both,ModeratelyRich,EarlyMorning,TwoHours,ModeratePlay,Secondary,Secondary,LessOften,Strong,GovernmentOwnedSchool,LessOften,MostOftenAvailab
2251950721,Male,3.50-4.00,Both,ModeratelyRich,MidNight,MoreThanTwoHours,LowInPlay,Graduate,Secondary,LessOften,VeryStrong,GoodPrivateSchool,Never,AlwaysAvailable,Very
225120305,Male,2.50-2.99,Both,ModeratelyRich,EarlyMorning,TwoHours,ModeratePlay,Secondary,Graduate,MostOften,Strong,GovernmentOwnedSchool,Never,AlwaysAvailable,Health
2251950834,Female,3.00-3.49,FatherOnly ,ModeratelyRich,EarlyMorning,OneHour,LowInPlay,Secondary,Secondary,Always,VeryStrong,GoodPrivateSchool,Never,MostOftenAvailable
2251111324,Male,3.50-4.00,Both,VeryRich,MidNight,LessThanOneHour,LowInPlay,Secondary,Secondary,MostOften,Strong,GovernmentOwnedSchool,Always,AlwaysAvailable,Healthy,S
2251930061,Female,2.50-2.99,MotherOnly,VeryRich,MidNight,TwoHours,ModeratePlay,Secondary,PostGraduate,MostOften,Strong,GovernmentOwnedSchool,Never,AlwaysAvailable,Hea
2251951104,Female,3.00-3.49,Both,LowIncomeEarner,Evenings,OneHour,ModeratePlay,Graduate,Primary,Never,Strong,GovernmentOwnedSchool,Never,LessOftenAvailable,VeryHealth
2251951078,Female,3.00-3.49,AnotherPerson,LowIncomeEarner,Evenings,OneHour,ModeratePlay,Secondary,Primary,Never,VeryStrong,GovernmentOwnedSchool,Never,LessOftenAvailal
2251940548,Female,2.50-2.99,Both,LowIncomeEarner,MidNight,TwoHours,LowInPlay,Secondary,Primary,Never,Strong,BoardingSchool,MostOften,MostOftenAvailable,Healthy,Bike,V
2251930054,Female,3.00-3.49,Both,LowIncomeEarner,EarlyMorning,MoreThanTwoHours,LowInPlay,Graduate,Secondary,MostOften,VeryWeak,GoodPrivateSchool,Never,AlwaysAvailable
225190631,Female,3.00-3.49,Both,ModeratelyRich,EarlyMorning,OneHour,LowInPlay,Secondary,Secondary,Always,Strong,GovernmentOwnedSchool,Never,AlwaysAvailable,VeryHealth
2251930515,Male,3.00-3.49,AnotherPerson,LowIncomeEarner,MidNight,MoreThanTwoHours,LowInPlay,Primary,Primary,Never,Strong,GovernmentOwnedSchool,Never,LessOftenAvailabl
2251960443,Female,3.00-3.49,Both,LowIncomeEarner,MidNight,TwoHours,LowInPlay,Secondary,Primary,MostOften,VeryStrong,GoodPrivateSchool,Never,AlwaysAvailable,Healthy,Wa
```

Search the web and Windows
2:59 PM
12/9/2021

Figure 3: Data in CSV format.

The data file (Figure 3) is converted to an ARFF files format in WEKA. First, the CSV file was opened and a few programming codes with the @relation was place on top of the note pad text editor. Then we declared all the variables and end the declarations with @data as shown in Figure 4. Finally, we saved the CSV file with the codes with a name that has an extension, ARFF. Figure 4 shows the screen shot of the programmed data in an ARFF format.



```
@relation PerformancePrediction

@attribute Matric numeric
@attribute Gender {Male,Female}
@attribute SFSGPA {0.00-1.99,2.00-2.49,2.50-2.99,3.00-3.49,3.50-4.00}
@attribute SLG {MotherOnly,FatherOnly,both,AnotherPerson}
@attribute PFS {VeryRich,ModeratelyRich,LowIncomeEarner}
@attribute PST {EarlyMorning,MidDay,Evenings,MidNight}
@attribute PSD {OneHour,TwoHours,MoreThanTwoHours,LessThanOneHour}
@attribute LOP {ExcessivePlay,ModeratePlay,LowInPlay,NoPlay}
@attribute FEB {Primary,Secondary,Graduate,Postgraduate}
@attribute MEB {Primary,Secondary,Graduate,Postgraduate}
@attribute PER {Never,LessOften,MostOften,Always}
@attribute FSA {VeryStrong,Strong,Weak,VeryWeak}
@attribute SST {GoodPrivateSchool,GovernmentOwnedSchool,ArmyDayGovernmentSchool,BoardingSchool}
@attribute PHL {Never,LessOften,MostOften,Always}
@attribute FA {NeverAvailable,LessOftenAvailable,MostOftenAvailable,AlwaysAvailable}
@attribute HS {VeryHealthy,Healthy,ModeratelyHealthy,OftenSick}
@attribute ATF {Walk,Bike,ThreeWheeler,ShuttleBus}
@attribute CC {VeryConducive,ACTight,Hot,NotGoodAtAll}
@attribute SSP {Good,VeryGood,Excellent,Fail}
@attribute LF {ReadilyAvailable,Available,SometimesAvailable,NotAvailable}

@data
225200136,Male,3.50-4.00,AnotherPerson,VeryRich,MidNight,MoreThanTwoHours,ModeratePlay,Graduate,Graduate,MostOften,VeryStrong,GoodPrivateSchool,LessOften,AlwaysAvaila
2251940166,Female,3.00-3.49,MotherOnly,ModeratelyRich,Evenings,TwoHours,ModeratePlay,Secondary,Secondary,Never,Strong,GovernmentOwnedSchool,LessOften,LessOftenAvailab
2251950934,Female,3.00-3.49,AnotherPerson,LowIncomeEarner,EarlyMorning,MoreThanTwoHours,ModeratePlay,Secondary,Primary,Never,Strong,GoodPrivateSchool,Never,LessOftenA
2252070167,Female,3.50-4.00,Both,LowIncomeEarner,EarlyMorning,LessThanOneHour,LowInPlay,Graduate,Secondary,Always,VeryWeak,GovernmentOwnedSchool,Never,MostOftenAvaila
2252050719,Male,3.50-4.00,Both, LowIncomeEarner,MidNight,TwoHours,LowInPlay,Secondary,Secondary,MostOften,Strong,GovernmentOwnedSchool,LessOften,LessOftenAvailable,He
2252070529,Male,3.00-3.49,Both,ModeratelyRich,Evenings,TwoHours,ModeratePlay,Secondary,Primary,LessOften,VeryStrong,GovernmentOwnedSchool,Never,AlwaysAvailable,VeryHe
225200363,Male,3.00-3.49,Both,ModeratelyRich,MidNight,MoreThanTwoHours,LowInPlay,Graduate,Secondary,LessOften,Weak,GoodPrivateSchool,Never,MostOftenAvailable,Healthy,
2252050923,Male,3.00-3.49,AnotherPerson,ModeratelyRich,MidNight,MoreThanTwoHours,ModeratePlay,Graduate,Secondary,Always,VeryStrong,BoardingSchool,Always,MostOftenAvai
2252050913,Female,3.00-3.49,Both,ModeratelyRich,MidNight,MoreThanTwoHours,ModeratePlay,Secondary,Primary,MostOften,Strong,GovernmentOwnedSchool,LessOften,MostOftenAva
2251950283,Male,3.00-3.49,Both,LowIncomeEarner,Evenings,TwoHours,ModeratePlay,Secondary,Secondary,Always,Strong,GovernmentOwnedSchool,Never,MostOftenAvailable,VeryHea
2251960019,Male,3.00-3.49,Both,ModeratelyRich,MidNight,OneHour,ModeratePlay,Graduate,Primary,Always,Weak,GoodPrivateSchool,Never,AlwaysAvailable,Healthy,Bike,Hot,Very
2251930060,Female,2.50-2.99,MotherOnly,LowIncomeEarner,MidNight,TwoHours,ModeratePlay,PostGraduate,Secondary,MostOften,Strong,GoodPrivateSchool,Always,AlwaysAvailable
```

Figure 4: Programmed ARFF file

4.2 Data Pre-processing

As it is common in data mining, before running tests on data instances, it is necessary to clean and prepare the data for use into the WEKA workbench. What is required is to convert string data into nominal data from the ARFF file. This was done based on the requirements constraints of the algorithms used. In addition, we looked at the relevance of the attributes to remove redundant, noisy, or irrelevant features. In the data, two attributes, students register number and their name were removed. WEKA has a functionality that enable missing values to be replaced. This was used to replace all missing values for attributes (see Figure 5). Replacing missing values places the distribution towards the mean value of the most frequent values for an attribute, and prevents the loss of information, which might potentially be useful for learning.

The Ranker method was used to select the top 10 attributes from those 20 attributes for this study. The attributes were ranked using select attributes functionality: an attribute evaluator and a search method object. The evaluator determines what method is used to assign a worth to each subset of attributes. The search method determines what style of search is performed. In this study. Attribute selection involves searching through all possible combinations of attributes in the data to find which subset of attributes works best for prediction.

A total of 169 records were taken for the analysis. Given the training set, the Naive Bayes algorithm first estimates the prior probability, P (Cj) for each class by counting how often each class occurs in the training data. For each attribute value, x_i was counted to determine $P(x_i)$. The probability, $P(x_i \mid C_j)$ was estimated by counting how often each value occurs in the class of the training data.

When classifying a target tuple, the conditional and prior probabilities generated from the training set were used to make the prediction. The $P(t_i \mid C_j)$ was estimated by (3) as;

$$P\left(t_i \middle| C_j\right) = \prod_{k=1}^{p} \left\{x_{ij} \middle| C_j\right\} \qquad (3)$$

In calculating $P(t_i)$ we estimated the likelihood that t_i is in each class. The probability that t_i is in a class, is the product of the conditional probabilities for each attribute value. The class with the highest probability was chosen for the tuple.

The present investigation used data mining as a tool with Naïve Bayes classification algorithm as a technique to design the student performance prediction model. Filtered feature selection technique was used to select the best subset of variables on the basis of the values of probabilities.

4.3 Dataset Analysis

Before setting the parameters for training and testing, the relevance of the attributes and the relationship of each attribute with the student's grade was determined. The relevance of each attribute and the relationship of the attributes with the Grade of the student was computed using spearman ranked correlation as shown in Figure 5 and Figure 6.

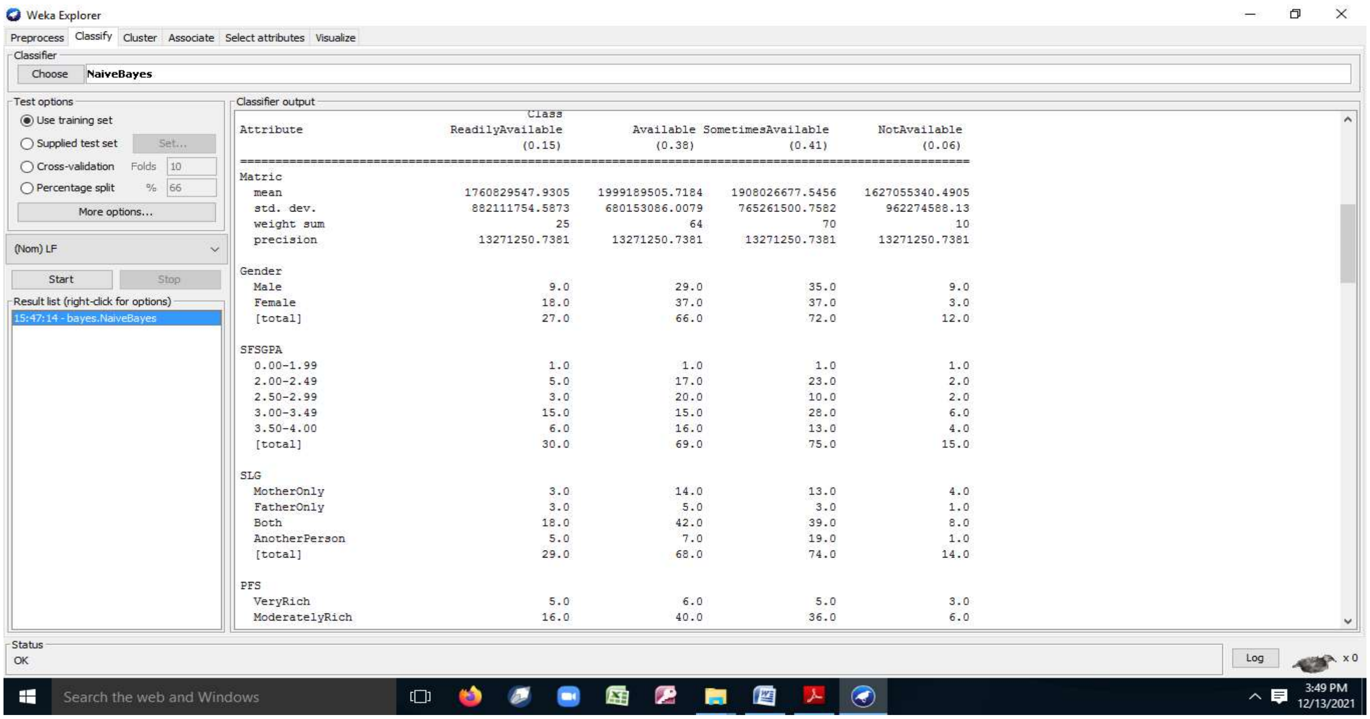


Figure 5: The output result of dataset analysis using WEKA

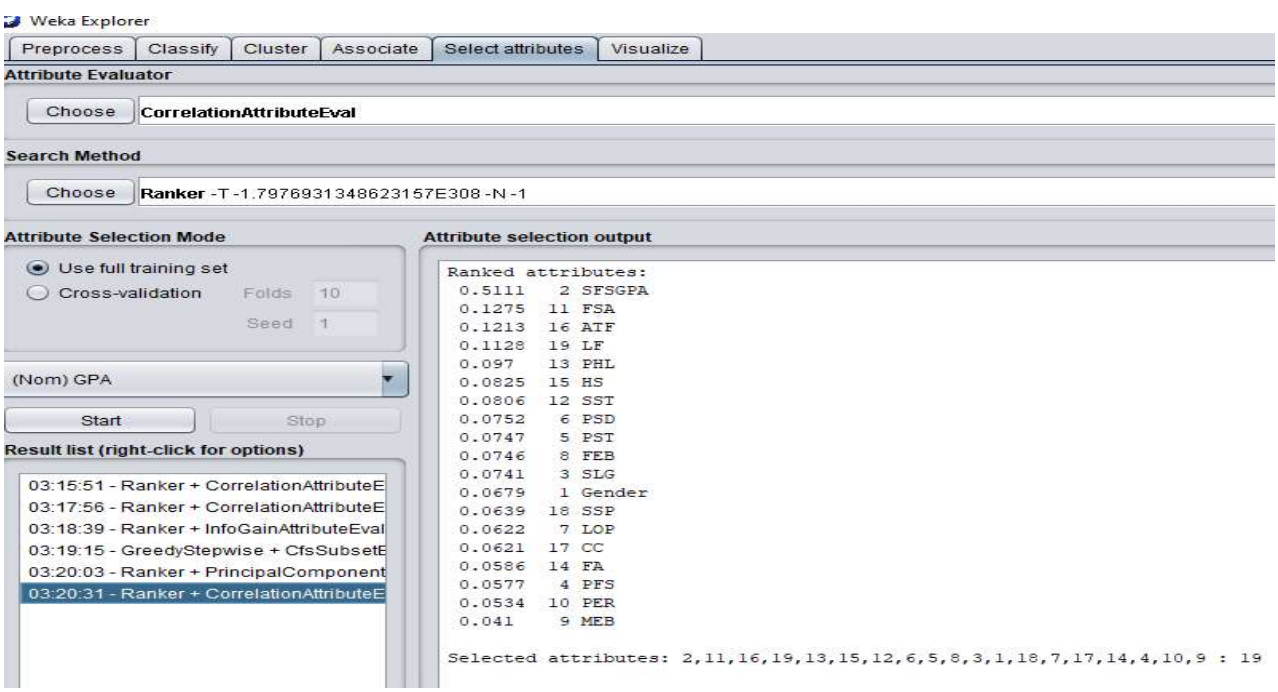


Figure 6: Variable importance

Figure 7: Plot of Variable/attribute importance

Figures 6 and Figure 7 show that SFSGPA has the highest relevance followed by FSA, GATF, LT and PHL. Other variables like HS, SST, PSD, PST, FEB, SLG, Gender, BSSP, LOP, CC, FA, PFS and PER have a very weak ranked correlation (<=0.1). MEB has a zero correlation.

4.4 Training and Testing

WEKA offers four test modes, which are used in training set, Supply test set, Cross-validation, and Percentages split. In this study, the cross-validation technique was selected for the training and testing. The cross-validation technique was set to three (3) folds validation in which the datasets are divided into 3 subsets. One of the 3 subsets is used as the testing data and others 2 subsets are put together to training the model. The training and testing data accuracy is average in 3 trails. The classifier output in Figure 8 shows the detailed accuracy by class and confusion matrix table.

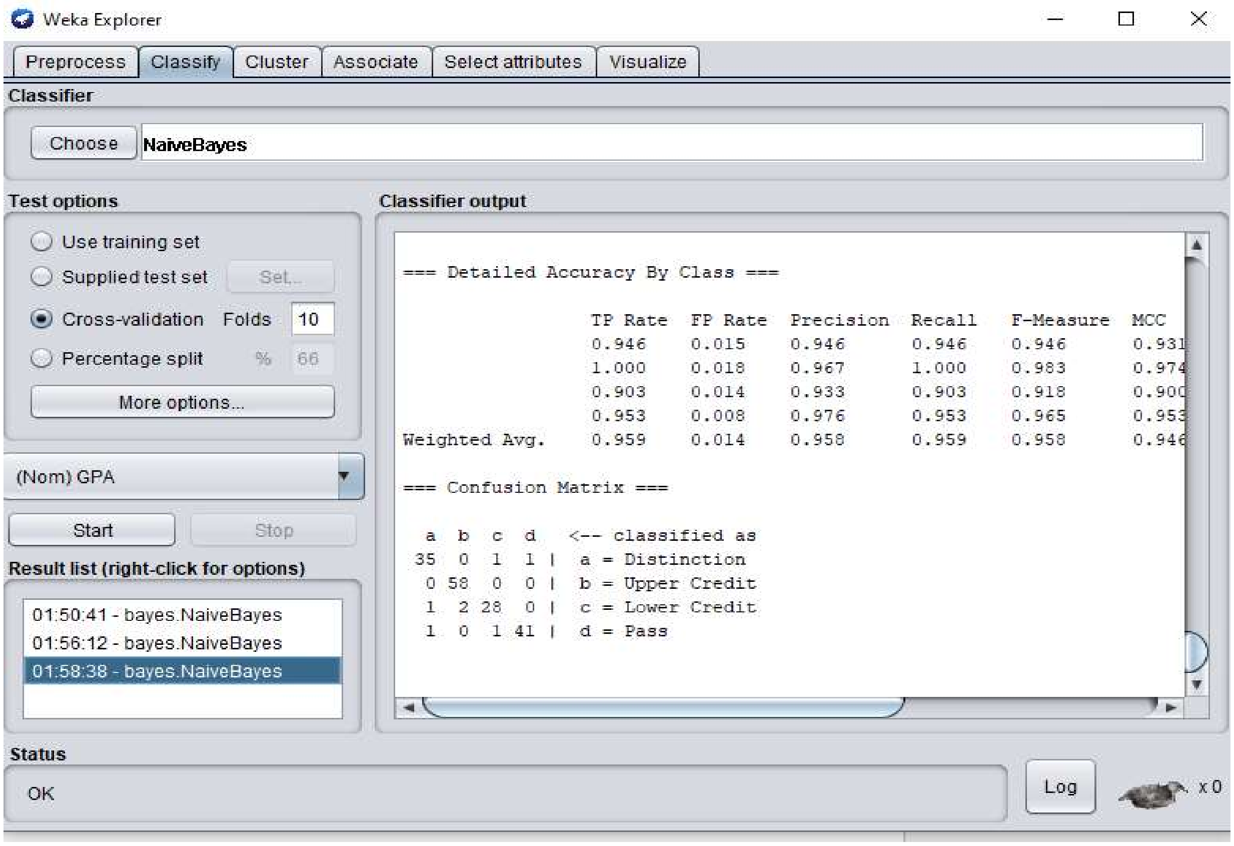


Figure 8: Model Settings for training

4.5 Result Analysis

The attributes were first analyzed for relevance, the analysis shows that SFSGPA has the highest relevance followed by FSA, GATF, LT and PHL. The analysis revealed that other variables like HS, SST, PSD, PST, FEB, SLG, Gender, BSSP, LOP, CC, FA, PFS and PER have a very weak ranked correlation (<=0.1), which implies that they may not significantly determine student's academic performance as shown in Figures 6 and 7.

One hundred and sixty-nine (169) students' records were used for the analysis. The model recorded 162 instances that were correctly classified which translates to approximately 96% accuracy while 7 instances were wrongly classified, which translates to approximately 4%. Figure 9, shows an excellent performance from the NBC model predicting the dataset correctly with 95.86% accuracy and incorrectly with 4.142%. The confusion matrix table at the bottom left corner of the screenshot in Figure 9 shows that for distinction class, 35 students were accurately classified while 2 students were misclassified. For Upper Credit class, 58 students were accurately classified, while there is no misclassified student. For Lower Credit class, 28 students were accurately classified while 3 students were misclassified. For Pass class, 41 students were accurately classified while 2 students were misclassified.

```
00:54:04 - bayes.NaiveBayes

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         162               95.858  %
Incorrectly Classified Instances         7                4.142  %
Kappa statistic                          0.9436
K&B Relative Info Score              14710.4919 %
K&B Information Score                  288.9555 bits      1.7098 bits/instance
Class complexity | order 0             331.589  bits      1.9621 bits/instance
Class complexity | scheme               43.3859 bits      0.2567 bits/instance
Complexity improvement     (Sf)        288.2031 bits      1.7053 bits/instance
Mean absolute error                      0.0678
Root mean squared error                  0.1352
Relative absolute error                 18.4029 %
Root relative squared error             31.5162 %
Total Number of Instances              169

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.946    0.015    0.946      0.946    0.946      0.931    0.999     0.996     Distinction
                 1.000    0.018    0.967      1.000    0.983      0.974    1.000     0.999     Upper Credit
                 0.903    0.014    0.933      0.903    0.918      0.900    0.995     0.978     Lower Credit
                 0.953    0.008    0.976      0.953    0.965      0.953    0.997     0.993     Pass
Weighted Avg.    0.959    0.014    0.958      0.959    0.958      0.946    0.998     0.993

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 35  0  1  1 |  a = Distinction
  0 58  0  0 |  b = Upper Credit
  1  2 28  0 |  c = Lower Credit
  1  0  1 41 |  d = Pass
```

Figure 9: Model Performance

5. CONCLUSION

The study has showed that Naïve Baye's model can efficiently be used to predict student academic performance for the degree of accuracy achieved. The analysis demonstrated that Naïve Baye's model is an ideal machine learning model where there is need to deal with numerous attributes. In this study Twenty (20) attributes were selected to evaluate students' academic performance based on academic and personal data collected. The input variables were collected through questionnaire method. The WEKA environment provided functionalities to determine which attribute that were selected based on their relevance in terms of its influence on the overall student academic performance. Since the effectiveness of predictive model depends on the volume of the training data, the cross-validation technique was chosen in the light of the size of the dataset that was available for the experiment. After testing some hypothesis, some of most influencing factors were identified and taken to predict the grades. It was found that Naïve Bayes algorithm performed very well therefore suitable for predicting student academic performance.

REFERENCES

Amjad A. S. (2016). Educational Data Mining & Students' Performance Prediction. (IJACSA) *International Journal of Advanced Computer Science and Applications, 7*(5), 2016. 212. www.ijacsa.thesai.org

Annisa, U. K. & Harwati, K. (2019). Educational Data Mining Techniques Approach to Predict Student's Performance. *International Journal of Information and Education Technology, 9*(2),

Aysha A., Sajid A., & Muhammad, G. K. (2018). A Comparative Study of Predicting Student's Performance by use of Data Mining Techniques. *American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS) (2018) 44*(1) 122-136

Usman, B., Adamu, R., & Salisu, S. (2020). Data Mining: Predicting of Student Performance Using Classification Technique. Int J. Information Processing and Comm. (IJIPC), 8(1), 92-101.

Danial H., Margus, P. & Yeongwook, Y. (2019). Mining Educational Data to Predict Students'Performance through Procrastination Behavior. MDPI

Ganorkar, S.S., Tiwari, N., Namdeo, V. (2021) Analysis and Prediction of Student Data Using Data Science: A Review. In: Zhang, YD., Senjyu, T., SO–IN C., Joshi, A. (eds) Smart Trends in Computing and Communications: *Proceedings of SmartCom 2020. Smart Innovation, Systems and Technologies, 182*. Springer, Singapore. https://doi.org/10.1007/978-981-15-5224-3_44

Hasan, R., Palaniappan, S., Mahmood, S., Abbas, A., Sarker, K. U., & Sattar, M. U. (2020). Predicting student performance in higher educational institutions using video learning analytics and data mining techniques. *Applied Sciences*, 10(11), 3894.

Kalpesh P. C., Riya A. S., Shreya S. J., & Rajeshwari J. B. (2017) Student Performance Prediction System using Data Mining Approach. *IJARCCE International Journal of Advanced Research in Computer and Communication Engineering ISO 3297:2007, 6*(3), March 2017. ISSN (Online) 2278-1021 ISSN (Print) 2319 5940. DOI10.17148/IJARCCE.2017.63195 833

Mohammed, H. S. & Nawzat, S. A. (2019). Classifying and Predicting Students' Performance using Improved Decision Tree C4.5 in Higher Education Institutes. *Journal of Computer Science, 3*(2)

Nawal A. Y., Rasha, G. M. H. & Somia B. M. (2017). Predicting Student Academic Performance in KSA using Data Mining Techniques. *Journal of Information Technology & Software Engineering, 7*(5). DOI: 10.4172/2165-7866.1000213

Rahila U., Teo S., Anuradha M. & Suriadi S. (2017). Predicting academic performance with process mining in learning analytics. *J. of Research in Innovative Teaching & Learning JRIT&L 10*(2).

Saheed, Y. K., Oladele, T. O., Akanni, A. O. & Ibrahim, W. M. (2018). Student Performance Prediction Based on Data Mining Classification Techniques. *Nigerian Journal of Technology (NIJOTECH),* 37(4) 1087 – 1091. Print ISSN: 0331-8443, Electronic ISSN: 2467-8821. www.nijotech.com http://dx.doi.org/ 10.4314/njt.v37i4.31

Sellappan, P., Salman M., Ali A., Kamal U. S. & Mian U. S. (2020). Predicting Student Performance in Higher Educational Institutions Using Video Learning Analytics and Data Mining Techniques. *(MDPI) Appl. Sci.* 2020, *10*(3) 894; doi:10.3390/app10113894 www.mdpi.com/journal/applsci

Sirwan M. A., & Ardalan H. A. (2019). Performance Analysis and Prediction Student Performance to build effective student Using Data Mining Techniques. *UHD Journal of Science and Technology 4(19)* DOI: 10.21928/uhdjst.v3n2y2019.pp10-15 https://www.researchgate.net/publication/334109444.